Published online 18 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W215–W222***
doi:10.1093/nar/gkr363

SDM—a server for predicting effects of mutations on protein stability and malfunction

Catherine L. Worth[1,2], Robert Preissner[2] and Tom L. Blundell[1,*]

[1]Biochemistry Department, University of Cambridge, Cambridge CB2 1GA, UK and [2]Institute for Physiology, Charité Universitätsmedizin, Lindenberger Weg 80, 13125 Berlin, Germany

Received February 18, 2011; Revised April 20, 2011; Accepted April 27, 2011

ABSTRACT

The sheer volume of non-synonymous single nucleotide polymorphisms that have been generated in recent years from projects such as the Human Genome Project, the HapMap Project and Genome-Wide Association Studies means that it is not possible to characterize all mutations experimentally on the gene products, i.e. elucidate the effects of mutations on protein structure and function. However, automatic methods that can predict the effects of mutations will allow a reduced set of mutations to be studied. Site Directed Mutator (SDM) is a statistical potential energy function that uses environment-specific amino-acid substitution frequencies within homologous protein families to calculate a stability score, which is analogous to the free energy difference between the wild-type and mutant protein. Here, we present a web server for SDM (http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php), which has obtained more than 10 000 submissions since being online in April 2008. To run SDM, users must upload a wild-type structure and the position and amino acid type of the mutation. The results returned include information about the local structural environment of the wild-type and mutant residues, a stability score prediction and prediction of disease association. Additionally, the wild-type and mutant structures are displayed in a Jmol applet with the relevant residues highlighted.

INTRODUCTION

Primarily hydrophobic interactions and a network of hydrogen bonds stabilize the folded state of a protein. However, a protein that is folded correctly is only marginally more stable than when it is unfolded, and mutations that affect a stabilizing interaction within a folded protein may lead to protein instability and malfunction. Where protein malfunction does occur and cannot be remediated by an alternative molecular pathway this may result in disease. For example, destabilizing mutations in phenylalanine hydroxylase lead to the metabolic disease, phenylketonuria (1). In fact, up to 80% of Mendelian disease-associated single mutations in protein coding regions are estimated to be caused by protein destabilization effects (2). However, a huge volume of single nucleotide polymorphisms (SNPs) has been generated in recent years from projects such as the Human Genome Project (3) and the HapMap Project (4) largely due to the availability of high-throughput array-based genotyping methods (5) and next generation sequencing platforms (6,7). Automatic methods that can predict the effect of mutations accurately will allow a reduced set of mutations to be characterized experimentally, saving time and money.

Various methods of predicting protein stability changes caused by mutation have been described and can be grouped into four main categories based on the strategy used in the calculation: (i) physical effective energy functions; (ii) empirical potential energy functions; (iii) machine learning methods; and (iv) statistical potential energy functions.

Physical potential energy functions (such as molecular mechanics approaches or Monte Carlo simulations) are probably the most accurate methods for predicting the effects of mutations on protein stability, however, they are currently only useful for testing small sets of mutants due to the large amount of time required to compute calculated $\Delta\Delta G$ values (8–12). The reliability of predictions is also complicated by the difficulties in sampling in the folded and unfolded states (12). Empirical potential approaches are fitted to experimental data using a set of weighted terms incorporating physical and statistical energy terms and structural descriptors (13,14). Machine learning methods include neural networks and support vector machines (SVMs) and use information about mutations, protein sequence and

*To whom correspondence should be addressed. Tel: +441223 333628; Fax: +441223 766002; Email: tom@cryst.bioc.cam.ac.uk

structural information to fit a non-linear function to experimental data (15–17). They are similar to empirical potential approaches in their use of experimental data to fit their function and in both cases, care must be taken that the function is not over-fitted to the training data set. Statistical potential energy approaches are derived from the statistical analysis of protein data such as substitution frequencies, distance potentials and amino acid environmental propensities (18–21). Other methods use a combination of the above strategies (22–24).

Site Directed Mutator (SDM) is a statistical potential energy function developed by Topham *et al.* (20) to predict the effect that SNPs will have on the stability of proteins. SDM uses environment-specific amino acid substitution frequencies within homologous protein families to calculate a stability score, which is analogous to the free energy difference between a wild-type and mutant protein. Blind testing on a set of 83 staphylococcal nuclease and 63 barnase mutants showed a correlation of 0.80 between the predicted stability changes and experimental data (20). The method performs comparably or better than other published methods in the task of classifying mutations as stabilizing or destabilizing (25). Additionally, SDM has much improved sensitivity in predicting stabilizing mutations compared to other published methods (five of the seven methods tested incorrectly classify >68% of the stabilizing mutations). When applied to the task of predicting disease-associated mutations, SDM had an accuracy of 61% (26). Therefore, SDM is a useful tool for guiding the design of site-directed mutagenesis experiments or for predicting whether a mutation will impact protein structure and have a role in disease. Here, we present a web server for SDM (http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php), which has not previously been published.

MATERIALS AND METHODS

Environment-specific substitution tables

SDM uses a set of conformationally constrained environment-specific substitution tables (ESSTs), the general methodology of which are described in (27,28). The tables were derived from 371 protein family sequence alignments from the HOMSTRAD database (29), consisting of 1357 structures and were built using a modified version of the program Makesub, which is able to handle sidechain hydrogen bond satisfaction (C. Topham, unpublished data). By defining the local structural environment of amino acid residues (secondary structure, solvent accessibility and formation of hydrogen bonds) distinct patterns of substitutions have been observed (30,31). Environment-specific substitution tables (ESSTs) store these substitution data quantitatively in the form of probabilities and therefore provide information about the existence of each amino acid in a particular environment and the probability of it being substituted by any other amino acid. Functional residues [as defined by Uniprot (32), the Catalytic Site Atlas (33) and Interpare (34)] were masked from substitution counts.

Definition of structural environment

The structural parameters that were used to define the local environment of amino acid residues are mainchain conformation, solvent accessibility and hydrogen-bonding class.

(i) Mainchain conformation and secondary structure: Nine classes of mainchain conformation were defined: residues were identified as belonging to either α-helix or β-sheet first and the remaining residues were classified as being *a, b, p, t, l, g* or *e* according to their mainchain φ-ψ torsion angles. The torsion angles and secondary structure assignments were calculated using the SSTRUC program (D. Smith, unpublished data).

(ii) Relative sidechain solvent accessibility: Three classes of relative sidechain solvent accessibility were defined based on the method of Lee and Richards (35). Residues with sidechain relative accessibilities of:
   (a) <17% were defined as inaccessible
   (b) 17–43% were defined as partially accessible
   (c) >43% were defined as accessible

These cut-offs were chosen based on an assessment of relative sidechain solvent accessibility values (36). The accessibility of each residue in a structure was calculated using the program PSA (A. Sali, unpublished data).

(iii) Hydrogen bonding: Two classes of hydrogen bonding were defined: residues were classed as either being satisfied in terms of their sidechain hydrogen bonding or not based on the criteria described by Worth and Blundell (37). Proteins were first protonated and the charge state of ionisable residues determined using the program, PROPKA (38). The program, HBOND (J. Overington, unpublished data), was used to identify hydrogen bonds defined by the criterion that the distance between donor and acceptor was <3.5Å except for interactions involving sulphur atoms where 4.0Å was used. Hydrogen bonds were then further filtered using the methodology described by Worth and Blundell (37).

These structural parameters gave a total of 54 local environments (nine mainchain × three solvent accessibility × two hydrogen bonding terms).

Prediction of protein stability changes caused by mutation

The algorithm underlying SDM was first described by Topham *et al.* (20). In this original work, two stability difference scores were calculated using either amino acid environmental substitution data (method I) or amino acid propensities (method II). Our subsequent analysis showed that updating the substitution and propensity data using additional protein families resulted in a better performance when the environment substitution data were used (data not shown). Therefore, SDM uses only method I to calculate protein stability changes caused by mutation. In addition, SDM now uses a far more comprehensive set of substitution data (ESSTs) compared to the original

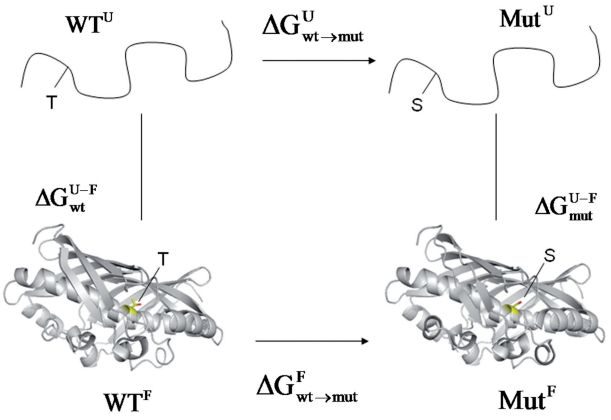

Figure 1. The thermodynamic cycle can be used to calculate protein stability changes between wild-type and mutant proteins.

publication (371 families compared to 131) and known functional sites are excluded from the substitution counts. Furthermore, the local structural environment parameter 'sidechain hydrogen bond (yes/no)' was modified to 'sidechain hydrogen-bonding satisfaction (satisfied/unsatisfied)' and this was shown to improve the stability score calculations (36).

By analogy to the folding-unfolding cycle in Figure 1, the algorithm uses ESSTs to calculate the difference in the stability scores of the folded and unfolded state for the wild-type and mutant protein structures:

$$\Delta\Delta s = \Delta s^{U}_{jk} - \Delta s^{F}_{jk} \quad (1)$$

The substitution data used for calculating the stability score are from families of homologous proteins, which have accepted multiple mutations during the course of their evolution. However, the effects of single substitutions are not often observed over the timescale of evolution e.g. cavity mutants. In order to compensate for this a disruption term is introduced for buried mutated residues. It is defined as the logarithmic function of the absolute value of the net change over the mutated position in the sidechain surface accessible area in an extended peptide Gly-X-Gly, relative to that for glycine. Therefore Equation (1) becomes:

$$\Delta\Delta s = \Delta s^{U}_{jk} - \Delta s^{F}_{jk} - \Delta s^{\mathrm{Disrupt}}_{jk} \quad (2)$$

ESSTs take into account the environment of only one of the two residues (wild-type or mutant), therefore it is necessary to consider not only the probability of replacement of the wild-type residue (R_j) in the wild-type environment (ε_{wt}) by a mutant residue type (r_k) in an undefined environment [$P(r_k/R_j, \varepsilon_{wt})$] but also the probability of replacement of the mutant residue type (R_k) in the mutant environment (ε_{mut}) by the wild-type residue (r_j) in an undefined environment [$P(r_j/R_k, \varepsilon_{mut})$].

In order to normalise the probabilities that are combined from different substitution tables, it is necessary to introduce a reference state. For the wild-type residue (R_j) in the wild-type environment a suitable reference state is the probability of it being conserved in that environment [$P(r_j/R_j, \varepsilon_{wt})$]. In an analogous way, for the mutant residue type (R_k) in the mutant environment, a suitable reference state is the probability of it being conserved in that environment [$P(r_k/R_k, \varepsilon_{mut})$].

The difference in stability scores for a mutation in the folded state is therefore calculated by:

$$\Delta s^{F}_{jk} = -\ln\left\{\frac{P(r_k/R_j,\varepsilon_{wt})}{P(r_j/R_j,\varepsilon_{wt})}\cdot\frac{P(r_k/R_k,\varepsilon_{mut})}{P(r_j/R_k,\varepsilon_{mut})}\right\} \quad (3)$$

The difference in stability scores in the unfolded state (Δs^{U}_{jk}) is also calculated using Equation (3) but uses an environmental substitution table derived from non-hydrogen bonded, surface exposed amino acid residues falling outside regions of regular secondary structure. The stability difference scores for the folded and unfolded state for the wild-type and mutant protein structures are then calculated using Equation (1).

Prediction of disease-association

From studying missense mutations for which the phenotypes are known, it is estimated that the stability margin that can be accommodated without any immediate effect on protein fitness is 1–3 kcal mol^{-1} (39–41). Studies of Ig-like proteins have shown that mutations that decrease the stability of these proteins by >2 kcal mol^{-1} result in severe disease phenotypes (42,43).

It may appear counter-intuitive that increased protein stability can lead to protein malfunction; however, protein flexibility is essential for enzyme catalysis. For instance, the increased stability of many thermophilic proteins is accompanied by loss of protein flexibility and reduced enzymatic activity at low temperatures (44–48). Furthermore, stabilizing mutations at catalytic site residues typically decrease activity and suggest that function often comes with a substantial penalty to stability (44,49–52). In addition, highly stable proteins are protease-resistant and therefore difficult to regulate—this is important to consider in systems such as cell signalling, where removing a signal is as important as its activation (53). A recent study showed that β-catenin accumulation is the most common aberration in parathyroid tumours of primary origin and that the S37A stabilizing mutation of CTNNB1 was found in 5.8% of the tumours (54). Another example of a stabilizing and damaging mutation is the Parkinson disease-associated A30P mutation, which stabilizes α-synuclein against proteasomal degradation triggered by haeme oxygenease-1 over-expression in human neuroblastoma cells (55). Hence, there is biological evidence that increased protein stability can lead to protein malfunction and hence disease.

In light of the studies mentioned in the previous two paragraphs, we have used a cut-off of 2 kcal mol^{-1} (stabilizing or destabilizing) for classifying mutations as leading to protein malfunction and possibly disease.

Mutant thermodynamic data sets

A subset of the data set used by Capriotti *et al.* (16) was used for initial benchmarking. This mutant data set was taken from the ProTherm database, which stores thermodynamic data for proteins and mutants (56). Our method requires knowledge of the local structural environment of wild-type and mutant residues in order to predict the effect of mutation on the stability of a protein. If the local environment is incorrectly defined e.g. the protein functions as a trimer but is defined in the crystallographic asymmetric unit as the protomer, this may affect our calculation. To remove the effect of such errors we used the Protein Interfaces, Surface and Assemblies (PISA) service to predict the oligomeric state of each of the proteins in the data set (57). Only those proteins predicted to be monomers were used. This data set is hereafter referred to as the *monomeric* set.

The validation data set used by Dehouck *et al.* (22) for benchmarking their method PoPMuSiC-2.0 was used for comparison of SDM's performance to other published stability change prediction algorithms. This data set comprises 350 mutations, none of which was included in any of the databases used to devise or test the seven methods tested by Dehouck *et al.* (22).

A set of 388 mutants (*S388*) with thermodynamic measurements conducted under physiological conditions was also used to test our method. The *S388* data set has been used to test other published methods and therefore allows us to perform a direct comparison of our method to them.

WEBSERVER

Input

SDM requires the 3D co-ordinates of the wild-type protein (in PDB format), the PDB chain identifier, the mutation position and the amino acid type of the mutation in one-letter code in order to calculate a stability score for mutant proteins. Users who have not already obtained a structure of their protein of interest may use the search boxes on the home page to do so. These search boxes allow a user to query the RCSB Protein Data Bank (www.pdb.org) (58) for their protein of interest, using protein name, description or amino acid sequence.

The wild-type structure may be submitted using one of two methods; the user can either upload the PDB file or enter the four-letter PDB code. NMR structures are accepted by SDM for input; however, users should note that it is only the first model in the PDB file, which is used for subsequent analysis.

SDM also requires a 3D structure of the mutant protein to perform its calculations. In this case, the user has the option of either uploading a mutant structure or using the program ANDANTE to build a model structure of the mutant (59). A requirement of SDM is that the wild-type and mutant structures span the same part of the polypeptide chain; therefore users must ensure that when they upload a mutant PDB structure that they fulfil this requirement.

The home page also provides a link to example output in order that users may view the type of output produced before running their job. Additionally, tutorials on usage are available for viewing using the link provided on the navigator bar.

Output

The results page is split into three sections. On the left-hand side the mutant information is displayed (wild-type and mutant amino acid types plus the position). Where ANDANTE was used to build a mutant structure, the PDB file is made available for download. The results returned include information about the local structural environment of the wild-type and mutant residues (the secondary structure, solvent accessibility and sidechain hydrogen bond satisfaction), a stability score prediction and prediction of disease association. As mentioned in the methods section, a cut-off of $2\,\mathrm{kcal\,mol^{-1}}$ is used to indicate whether a mutation is likely to be disease-associated or not. However, mutations that do not reach this cut-off may still lead to protein malfunction and disease if they affect binding sites. A statement indicating this issue is therefore displayed and the links page lists resources that can be used to assess whether a residue is involved in binding.

In the middle portion of the results page, the wild-type and mutant structures are displayed using the Jmol structure viewer (Jmol: an open-source Java viewer for chemical structures in 3D http://www.jmol.org/) with the relevant residues highlighted. The user may control the display of these structures using the menu buttons on the right-hand side.

An example of the type of output produced by SDM is shown in Figure 2. A particular advantage of the predictions provided by SDM over other published methods is the indication of the local structural environment of wild-type and mutant residues and the fact that the user may view the 3D structural context of the residues. This allows users to identify possible molecular mechanisms that underlie predicted stability changes for example, loss of hydrogen bonds to the protein backbone.

VALIDATION

SDM has previously been validated using a set of ~230 mutants and was shown to have an accuracy of 74% in predicting the sign of stability change and a linear correlation coefficient of 0.60 between predicted and observed $\Delta\Delta G$ values (25). Removal of one outlying data point increased the linear correlation coefficient to 0.66. Analysis of the performance of SDM in predicting the sign of stability change in comparison to eight other published methods demonstrated that SDM performs comparably or better than the other methods.

Since the benchmarking detailed above was carried out, SDM has been modified so that the definition of sidechain hydrogen bonding has been changed from yes or no to satisfied or unsatisfied. Furthermore, functional residues have been masked from the substitution counts used to generate the ESSTs. We tested the improvement that

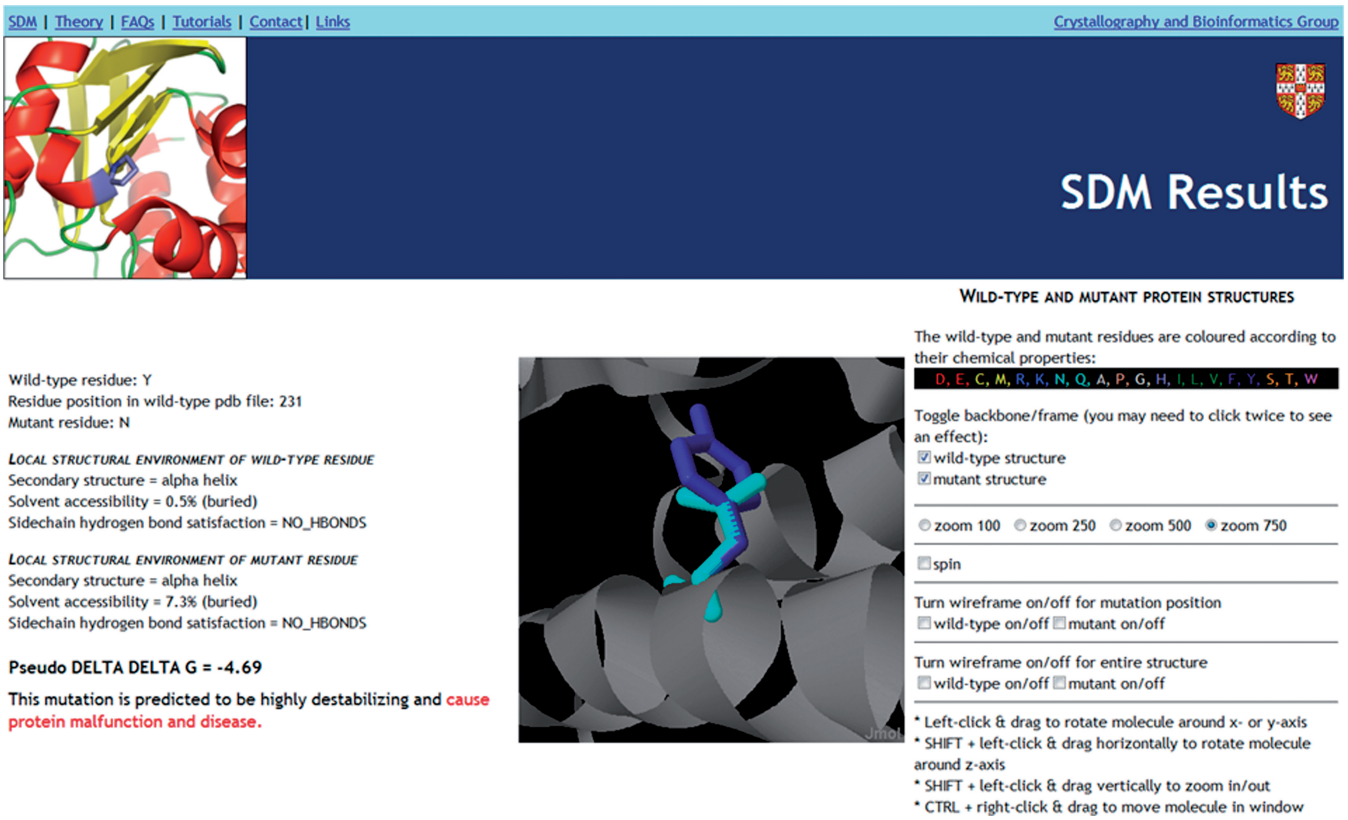

Figure 2. Screenshot of SDM analysis results for the example of mutation Y231N in Dystrophin (PDB code 1DXX, chain A). On the left hand side information about the wild-type and mutant residue is displayed such as the secondary structure, solvent accessibility and hydrogen bonds formed by the sidechain. Underneath this information is the predicted effect on protein stability. In this case, SDM predicts that the mutation is highly destabilizing and disease-associated. In fact, this mutation is associated with muscular dystrophy and has been shown to decrease protein stability (73). In the middle, the structural context of the wild-type and mutant amino acids are shown in the Jmol applet with the residues coloured according to their chemical properties (key displayed on right hand side). Using the menus on the right hand side the user can manipulate the Jmol applet and control what is shown.

Table 1. Comparison of the performance of SDM using different sets of ESSTs and the monomeric data set

Parameters used to generate ESSTs			Accuracy (%)	R^a	σ (kcal/mol)
Protein families	Hydrogen bonding term	Masking of functional residues			
113	Original	No	73	0.51	1.82
371	Original	Yes	73	0.56	1.61
371	Satisfied	No	73	0.56	1.73
371	Satisfied	Yes	71	0.58	1.74

[a]Pearson product-moment correlation coefficient.

these changes made to SDM's predictions using the 855 mutants in the *monomeric* data set. The additional families used to generate the ESSTs, masking functional residues and incorporation of the hydrogen bond satisfaction term improved the correlation coefficient between predicted stability changes and experimental measurements from 0.51 to 0.58 (Table 1).

The statistical potential-based method, PoPMuSiC-2.0 was recently reported and achieved a correlation of 0.63 between measured and predicted stability changes (22). The predictive power of the method was shown to be significantly higher than that of other programs described in the literature. In order to compare the predictive power of SDM to PoPMuSiC-2.0 and the other tested methods, we used the same data set of 350 mutants. After the PoPMuSiC algorithms, SDM has the highest linear correlation between predicted and measured $\Delta\Delta G$ values (Table 2). It also has the benefit of making predictions for the entire data set of 350 mutants. It is encouraging that the performance of SDM is improved when considering only highly stabilizing or destabilizing mutations—the correlation coefficient increases from 0.52 to 0.63 (Table 2).

The vast majority of published methods for predicting the effects of mutations on protein stability are based on machine learning (ML). These are first trained on a data set of mutations. Many of these ML methods report high correlations with experimental data sets [e.g. CUPSAT $R = 0.87$ (21) and IMutant2.0 $R = 0.71$ (60)]. However, when tested later in blind tests, these correlations drop drastically [e.g. CUPSAT $R = 0.37$ and IMutant-2.0 $R = 0.29$ (22)]. This reduction in prediction

Table 2. Comparison of the performance of different prediction methods

Method	No. of predictions[b]	Complete set (350/309/87 mutants)[a]	
		R	σ (kcal/mol)
Automute[c]	315	0.46 / 0.45 / 0.45	1.43 / 1.46 / 1.99
CUPSAT[c]	346	0.37 / 0.35 / 0.50	1.91 / 1.96 / 2.14
Dmutant[c]	350	0.48 / 0.47 / 0.57	1.81 / 1.87 / 2.31
Eris[c]	334	0.35 / 0.34 / 0.49	4.12 / 4.28 / 3.91
I-mutant-2.0[c]	346	0.29 / 0.27 / 0.27	1.65 / 1.69 / 2.39
PoPMuSiC-1.0[c]	350	0.62 / 0.63 / 0.70	1.24 / 1.25 / 1.66
PoPMuSiC-2.0[c]	350	0.67 / 0.67 / 0.71	1.16 / 1.19 / 1.67
SDM	350	0.52 / 0.53 / 0.63	1.80 / 1.81 / 2.11

[a]Three values are given per column. The first corresponds to the whole validation set of 350 mutants with the unavailable $\Delta\Delta G$ predictions set to 0.0 kcal/mol. The second corresponds to the 309 mutants for which a $\Delta\Delta G$ prediction is available for all predictors. The third corresponds to 87 mutants for which the experimental $\Delta\Delta G$ value causes >2 kcal mol^{-1} change and for which a $\Delta\Delta G$ prediction is available for all predictors.
[b]350 mutations were tested with each method. However, some servers failed to compute the ΔΔG prediction for all mutants, resulting in predictions for less than the full number.
[c]Data taken from (22).

Table 3. Comparison of the performance of different prediction methods

Method	MCC	Accuracy	Sens. (+)	Spec. (+)	Sens. (−)	Spec. (−)
Automute S1227[a]	0.31	0.87	0.36	0.42	0.94	0.92
FOLDX[b]	0.25	0.75	0.56	0.26	0.78	0.93
DFIRE[b]	0.11	0.68	0.44	0.18	0.71	0.90
PoPMuSiC-1.0[b]	0.20	0.85	0.25	0.33	0.93	0.90
PoPMuSiC-2.0	0.32	0.86	0.35	0.44	0.94	0.91
NeuralNet[b]	0.25	0.87	0.21	0.44	0.96	0.90
MuPro SO[c]	0.26	0.86	0.30	0.40	0.94	0.90
MuPro TO[c]	0.28	0.86	0.31	0.42	0.94	0.91
MuPro ST[c]	0.27	0.86	0.31	0.40	0.93	0.91
MuX-S[d]	0.39	0.88	0.29	0.67	0.94	0.91
MuX-48[c]	0.39	0.89	0.29	0.67	0.98	0.91
SDM	0.28	0.71	0.70	0.24	0.71	0.94

[a]Data taken from Masso and Vaisman (24).
[b]Data taken from Capriotti *et al.* (16).
[c]Data taken from Cheng *et al.* (17).
[d]Data taken from Kang *et al.* (74).

performance may be due to over-fitting to available data sets. The problem of decreasing performance of ML methods using blind-data sets was also observed by two independent assessments of the performance of protein stability predictors (61,62). SDM is not a ML method, but rather a statistical method based on observed amino acid substitutions that have occurred during divergent protein evolution. Therefore, it does not suffer from the problem of over-fitting, as demonstrated by the similar correlation coefficients obtained using the *monomeric* data set and the PoPMuSiC-2.0 validation data set. The problem of over-fitting is an important point to consider if methods are to be used to help successfully design mutagenesis experiments.

Table 3 shows the results of testing the *S388* data set. These results show the performance of methods in predicting the sign of stability change i.e. whether a mutation is stabilizing or destabilizing. Many of the methods have accuracies of over 80%, which is impressive. However, if we examine the ability of the methods to predict stabilizing and destabilizing mutations another picture emerges; they tend to be very good at predicting destabilizing mutations but much worse at predicting stabilizing mutations. SDM however has a more balanced sensitivity in predicting both types of mutations, although the specificity of predicting destabilizing mutations is far better than that of predicting stabilizing mutations. Most mutations are destabilizing and this is reflected in the mutant thermodynamic data sets used for developing and testing such methods. Methods that assign all of the samples to the majority class (destabilizing mutations) will have high accuracy even though the performance is poor for the minority class (stabilizing mutations). This trend is observed for most of the methods reported in Table 3. It is possible that some of the results in Table 3 are biased by some over-fitting to the training data sets used in developing the methods.

When applied to the task of predicting disease-associated mutations, SDM had an accuracy of 61% (26), only 3% less than the accuracy achieved by the program Sorting Intolerant from Tolerant (SIFT) (63). Of course, it is unsurprising that SIFT obtains a higher accuracy than SDM as SDM is able to distinguish disease-associations only for those mutations that perturb protein structure and not those that directly affect catalytic residues, binding sites etc. Mutations that cause protein malfunction by affecting the functional residues of a protein (active sites or protein–protein interaction sites) or by altering post-translational modifications will not be identified as damaging by SDM. Therefore, to obtain a more accurate prediction of whether an nsSNP is associated with disease, these other effects should also be taken into account. We previously demonstrated that when SDM's predictions were combined with predictions of functional sites using Crescendo (64) and known functional sites, this combined approach has a comparable accuracy to the other methods tested but has the benefit of a much lower false-positive rate, therefore providing a high-quality set of predictions (26).

SUMMARY

The SDM server provides users with a fast and accurate means of assessing the impact that a mutation will have on protein structure and stability. It provides a 3D view of the wild-type and mutant residues, allowing users to inspect the structural context of the sidechains. SDM is a useful tool for identifying possible disease associations and has been applied to the task of predicting deleterious nsSNPs at the genome scale (25,26,65) and also for generating new hypotheses regarding: (i) the molecular aetiology of renal cell carcinoma and pheochromocytoma in the cancer syndrome, von Hippel-Lindau disease (66); (ii) the structural effects of mutations in thyroid

stimulating hormone receptor that are associated with congenital non-goitrous hypothyroidism (67); and (iii) tumour risk associated with mutations in succinate dehydrogenase D (68). It has also been used in the analysis of mutations in the autoimmune regulator protein (69), mixed lineage kinase 3 (70), the adaptor protein MyD88 adaptor-like (71) and breast cancer susceptibility gene 1 (72).

FUNDING

This work was supported by the Biotechnology and Biological Sciences Research Council (research studentship to C.L.W.) and a Wellcome Trust Programme Grant (to T.L.B.). Funding for open access charge: Wellcome Trust Programme Grant.

Conflict of interest statement. None declared.

REFERENCES

1. Bjorgo,E., Knappskog,P.M., Martinez,A., Stevens,R.C. and Flatmark,T. (1998) Partial characterization and three-dimensional-structural localization of eight mutations in exon 7 of the human phenylalanine hydroxylase gene associated with phenylketonuria. *Eur. J. Biochem.*, **257**, 1–10.
2. Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.
3. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
4. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
5. Gunderson,K.L., Steemers,F.J., Ren,H., Ng,P., Zhou,L., Tsan,C., Chang,W., Bullis,D., Musmacker,J., King,C. *et al.* (2006) Whole-genome genotyping. *Methods Enzymol.*, **410**, 359–376.
6. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
7. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.J., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
8. Bash,P.A., Singh,U.C., Langridge,R. and Kollman,P.A. (1987) Free energy calculations by computer simulation. *Science*, **236**, 564–568.
9. Funahashi,J., Sugita,Y., Kitao,A. and Yutani,K. (2003) How can free energy component analysis explain the difference in protein stability caused by amino acid substitutions? Effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme. *Protein Eng.*, **16**, 665–671.
10. Kollman,P.A., Massova,I., Reyes,C., Kuhn,B., Huo,S., Chong,L., Lee,M., Lee,T., Duan,Y., Wang,W. *et al.* (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, **33**, 889–897.
11. Park,H. and Lee,S. (2005) Prediction of the mutation-induced change in thermodynamic stabilities of membrane proteins from free energy simulations. *Biophys. Chem.*, **114**, 191–197.
12. Shi,Y.Y., Mark,A.E., Wang,C.X., Huang,F., Berendsen,H.J. and van Gunsteren,W.F. (1993) Can the stability of protein mutants be predicted by free energy calculations? *Protein Eng.*, **6**, 289–295.
13. Bordner,A.J. and Abagyan,R.A. (2004) Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins*, **57**, 400–413.
14. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
15. Capriotti,E., Fariselli,P., Calabrese,R. and Casadio,R. (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21(Suppl. 2)**, ii54–ii58.
16. Capriotti,E., Fariselli,P. and Casadio,R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20(Suppl. 1)**, i63–i68.
17. Cheng,J., Randall,A. and Baldi,P. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.
18. Gilis,D. and Rooman,M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**, 276–290.
19. Saraboji,K., Gromiha,M.M. and Ponnuswamy,M.N. (2006) Average assignment method for predicting the stability of protein mutants. *Biopolymers*, **82**, 80–92.
20. Topham,C.M., Srinivasan,N. and Blundell,T.L. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**, 7–21.
21. Parthiban,V., Gromiha,M.M. and Schomburg,D. (2006) CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.*, **34**, W239–W242.
22. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
23. Yin,S., Ding,F. and Dokholyan,N.V. (2007) Modeling backbone flexibility improves protein stability estimation. *Structure*, **15**, 1567–1576.
24. Masso,M. and Vaisman,II. (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, **24**, 2002–2009.
25. Worth,C.L., Burke,D.F. and Blundell,T.L. (2007) Estimating the effects of single nucleotide polymorphisms on protein structure: how good are we at identifying likely disease associated mutations? *Proceedings of "Molecular Interactions - Bringing Chemistry to Life"*, 11–26.
26. Worth,C.L., Bickerton,G.R., Schreyer,A., Forman,J.R., Cheng,T.M., Lee,S., Gong,S., Burke,D.F. and Blundell,T.L. (2007) A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *J. Bioinform. Comput. Biol.*, **5**, 1297–1318.
27. Overington,J., Donnelly,D., Johnson,M.S., Sali,A. and Blundell,T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.
28. Topham,C.M., McLeod,A., Eisenmenger,F., Overington,J.P., Johnson,M.S. and Blundell,T.L. (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.*, **229**, 194–220.
29. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
30. Blundell,T.L., Cooper,J., Donnelly,D., Driessen,H., Edwards,Y., Eisenmenger,F., Frazao,C., Johnson,M., Niefind,K., Newman,M. *et al.* (1991) Patterns of sequence variation in families of homologous proteins. In Jornvall/Hoog/Gustavsson. (ed.), *Methods in Protein Sequence Analysis*. Birkhauser Verlag AG, Basel, pp. 373–385.
31. Overington,J., Johnson,M.S., Sali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.*, **241**, 132–145.
32. Consortium,U. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
33. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

34. Gong,S., Park,C., Choi,H., Ko,J., Jang,I., Lee,J., Bolser,D.M., Oh,D., Kim,D.S. and Bhak,J. (2005) A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **6**, 207.
35. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
36. Worth,C.L. (2008) The role of amino acid sidechains in protein stability. Ph.D Thesis. University of Cambridge, Cambridge.
37. Worth,C.L. and Blundell,T.L. (2009) Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. *Proteins*, **75**, 413–429.
38. Li,H., Robertson,A.D. and Jensen,J.H. (2005) Very fast empirical prediction and rationalization of protein pKa values. *Proteins*, **61**, 704–721.
39. Calloni,G., Zoffoli,S., Stefani,M., Dobson,C.M. and Chiti,F. (2005) Investigating the effects of mutations on protein aggregation in the cell. *J. Biol. Chem.*, **280**, 10607–10613.
40. Mayer,S., Rudiger,S., Ang,H.C., Joerger,A.C. and Fersht,A.R. (2007) Correlation of levels of folded recombinant p53 in escherichia coli with thermodynamic stability in vitro. *J. Mol. Biol.*, **372**, 268–276.
41. Tokuriki,N. and Tawfik,D.S. (2009) Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.*, **19**, 596–604.
42. Lindberg,M.J., Bystrom,R., Boknas,N., Andersen,P.M. and Oliveberg,M. (2005) Systematically perturbed folding patterns of amyotrophic lateral sclerosis (ALS)-associated SOD1 mutants. *Proc. Natl Acad. Sci. USA*, **102**, 9754–9759.
43. Randles,L.G., Lappalainen,I., Fowler,S.B., Moore,B., Hamill,S.J. and Clarke,J. (2006) Using model proteins to quantify the effects of pathogenic mutations in Ig-like proteins. *J. Biol. Chem.*, **281**, 24216–24226.
44. Counago,R., Wilson,C.J., Pena,M.I., Wittung-Stafshede,P. and Shamoo,Y. (2008) An adaptive mutation in adenylate kinase that increases organismal fitness is linked to stability-activity trade-offs. *Protein Eng. Des. Sel.*, **21**, 19–27.
45. Jaenicke,R. (1991) Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem.*, **202**, 715–728.
46. Somero,G.N. (1995) Proteins and temperature. *Annu. Rev. Physiol.*, **57**, 43–68.
47. Wolf-Watz,M., Thai,V., Henzler-Wildman,K., Hadjipavlou,G., Eisenmesser,E.Z. and Kern,D. (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.*, **11**, 945–949.
48. Zavodszky,P., Kardos,J., Svingor. and Petsko,G.A. (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl Acad. Sci. USA*, **95**, 7406–7411.
49. Beadle,B.M. and Shoichet,B.K. (2002) Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.*, **321**, 285–296.
50. Meiering,E.M., Serrano,L. and Fersht,A.R. (1992) Effect of active site residues in barnase on activity and stability. *J. Mol. Biol.*, **225**, 585–589.
51. Mukaiyama,A., Haruki,M., Ota,M., Koga,Y., Takano,K. and Kanaya,S. (2006) A hyperthermophilic protein acquires function at the cost of stability. *Biochemistry*, **45**, 12673–12679.
52. Yutani,K., Ogasahara,K., Tsujita,T. and Sugino,Y. (1987) Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. *Proc. Natl Acad. Sci. USA*, **84**, 4441–4444.
53. DePristo,M.A., Weinreich,D.M. and Hartl,D.L. (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.*, **6**, 678–687.
54. Bjorklund,P., Lindberg,D., Akerstrom,G. and Westin,G. (2008) Stabilizing mutation of CTNNB1/beta-catenin and protein accumulation analyzed in a large series of parathyroid tumors of Swedish patients. *Mol. Cancer*, **7**, 53.
55. Song,W., Patel,A., Qureshi,H.Y., Han,D., Schipper,H.M. and Paudel,H.K. (2009) The Parkinson disease-associated A30P mutation stabilizes alpha-synuclein against proteasomal degradation triggered by heme oxygenase-1 over-expression in human neuroblastoma cells. *J. Neurochem.*, **110**, 719–733.
56. Gromiha,M.M. and Sarai,A. (2010) Thermodynamic database for proteins: features and applications. *Methods Mol. Biol.*, **609**, 97–112.
57. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
58. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
59. Smith,R.E., Lovell,S.C., Burke,D.F., Montalvao,R.W. and Blundell,T.L. (2007) Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. *Bioinformatics*, **23**, 1099–1105.
60. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
61. Khan,S. and Vihinen,M. (2010) Performance of protein stability predictors. *Hum. Mutat.*, **31**, 675–684.
62. Potapov,V., Cohen,M. and Schreiber,G. (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng. Des. Sel.*, **22**, 553–560.
63. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
64. Chelliah,V., Chen,L., Blundell,T.L. and Lovell,S.C. (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.*, **342**, 1487–1504.
65. Burke,D.F., Worth,C.L., Priego,E.M., Cheng,T., Smink,L.J., Todd,J.A. and Blundell,T.L. (2007) Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, **8**, 301.
66. Forman,J.R., Worth,C.L., Bickerton,G.R., Eisen,T.G. and Blundell,T.L. (2009) Structural bioinformatics mutation analysis reveals genotype-phenotype correlations in von Hippel-Lindau disease and suggests molecular mechanisms of tumorigenesis. *Proteins*, **77**, 84–96.
67. Cangul,H., Morgan,N.V., Forman,J.R., Saglam,H., Aycan,Z., Yakut,T., Gulten,T., Tarim,O., Bober,E., Cesur,Y. *et al.* (2010) Novel TSHR mutations in consanguineous families with congenital nongoitrous hypothyroidism. *Clin. Endocrinol.*, **73**, 671–677.
68. Ricketts,C.J., Forman,J.R., Rattenberry,E., Bradshaw,N., Lalloo,F., Izatt,L., Cole,T.R., Armstrong,R., Kumar,V.K., Morrison,P.J. *et al.* (2010) Tumor risks and genotype-phenotype-proteotype analysis in 358 patients with germline mutations in SDHB and SDHD. *Hum. Mutat.*, **31**, 41–51.
69. Ferguson,B.J., Alexander,C., Rossi,S.W., Liiv,I., Rebane,A., Worth,C.L., Wong,J., Laan,M., Peterson,P., Jenkinson,E.J. *et al.* (2008) AIRE's CARD revealed, a new structure for central tolerance provokes transcriptional plasticity. *J. Biol. Chem.*, **283**, 1723–1731.
70. Velho,S., Oliveira,C., Paredes,J., Sousa,S., Leite,M., Matos,P., Milanezi,F., Ribeiro,A.S., Mendes,N., Licastro,D. *et al.* (2010) Mixed lineage kinase 3 gene mutations in mismatch repair deficient gastrointestinal tumours. *Hum. Mol. Genet.*, **19**, 697–706.
71. Nagpal,K., Plantinga,T.S., Wong,J., Monks,B.G., Gay,N.J., Netea,M.G., Fitzgerald,K.A. and Golenbock,D.T. (2009) A TIR domain variant of MyD88 adapter-like (Mal)/TIRAP results in loss of MyD88 binding and reduced TLR2/TLR4 signaling. *J. Biol. Chem.*, **284**, 25742–25748.
72. Rowling,P.J., Cook,R. and Itzhaki,L.S. (2010) Toward classification of BRCA1 missense variants using a biophysical approach. *J. Biol. Chem.*, **285**, 20080–20087.
73. Singh,S.M., Kongari,N., Cabello-Villegas,J. and Mallela,K.M. (2010) Missense mutations in dystrophin that trigger muscular dystrophy decrease protein stability and lead to cross-beta aggregates. *Proc. Natl Acad. Sci. USA*, **107**, 15069–15074.
74. Kang,S., Chen,G. and Xiao,G. (2009) Robust prediction of mutation-induced protein stability change by property encoding of amino acids. *Protein Eng. Des. Sel.*, **22**, 75–83.